Comments on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination

Peter Rousseeuw and Wannes Van den Bossche
Department of Mathematics, KU Leuven, Belgium

December 23, 2014

The paper by Agostinelli, Leung, Yohai and Zamar (henceforth ALYZ) marks an important step forward in that it proposes an estimator that is specifically designed to tackle the combination of cellwise (ICM) and casewise (THCM) outliers. As it was argued in the fast growing literature on cellwise contamination (e.g. Alqallaf et al. 2009, Van Aelst et al. 2012, Öllerer et al. 2013, Farcomeni 2014), cellwise outliers form a serious real-life problem which needs to be addressed. And whereas cellwise contamination is difficult enough by itself, the realistic situation in which both cellwise and casewise outliers may occur in the data is even harder. The authors make a point that the methods developed so far can handle one type of outliers or the other, but not yet both.

The 2SGS method proposed by ALYZ starts with an initial step in which coordinatewise outliers are flagged (relative to robust univariate estimates of location and scale), and then set to NA. But the real workhorse of 2SGS is the second step, the generalized S-estimator (GSE) of Danilov et al. (2012) which provides a robust estimate of multivariate location and scatter in the presence of missing values.

As far as we know 2SGS is currently the best available method for dealing with both cellwise and casewise outliers, as seen in the simulations provided by ALYZ as well as our own. But we think there remains room for improvement on two fronts, so further research would be welcome.

The first aspect that seems too limiting is that the initial step flags cellwise outliers based solely on their value for that single variable, so in this step no correlation structure is taken into account yet. On the other hand, it is well-known that a point can be outlying without any of its coordinates being outlying.

Table 1: Computation time in seconds

dimension	2SGS	HSD	SnipEM	CovScores	DetMcdScores
10	0.06	0.38	1.07	0.002	0.02
20	0.45	1.02	3.32	0.003	0.09
30	1.94	2.21	9.01	0.008	0.27
40	5.95	6.14	15.56	0.012	0.64
50	14.06	10.96	34.08	0.015	1.26

It would thus be better if the first step could already use some information about the correlation structure, which 2SGS considers in the second step.

In connection with this we drew histograms of the correlations between the uncontaminated variables in the simulation setup of ALYZ, and found that most correlations were rather small, especially with increasing dimension p. So we added a different setup in which the correlation matrix $\boldsymbol{C}$ between the uncontaminated variables has entries

$$\boldsymbol{C}_{jk} = \rho^{|j-k|} \tag{1}$$

with $0 < \rho < 1$. The casewise outliers were generated in the same way as in ALYZ, in the direction of the last eigenvector. But whereas ALYZ generated cellwise outliers by replacing cells by the value k, we put in the values k and $-k$ at random to compensate for the fact that the correlations in (1) are all positive.

The second aspect is computational complexity. The first step of 2SGS is fast but the second step isn't, due to the many iterations in the GSE algorithm. This hinders scalability of the method, as our Table 1 illustrates that the computation time goes up rapidly with the dimension p. We do not know whether this complexity is intrinsic but it seems worthwhile to search for faster approaches. The problem will be to combine speed with robustness.

One such simple but less robust approach is to replace each cell x_{ij} by

$$\mathrm{MAD}_j \; \Phi^{-1}(\frac{R(x_{ij})}{n+1}) \tag{2}$$

where MAD_j is the normalized median absolute deviation of variable j and $R(x_{ij})$ is the rank of x_{ij} in variable j. A trivial approach we will call *CovScores* is then to compute the classical covariance matrix of (2), as was done in one of the initial estimators of the deterministic minimum covariance determinant (DetMCD) algorithm of Hubert et al. (2012). This of course corresponds to computing the normal scores correlation matrix (also called the Van der Waerden rank correlation matrix). The same approach was taken by Öllerer and Croux (2014), who

used the resulting covariance matrix as input to the GLASSO algorithm of Friedman et al. (2008) to obtain a sparse estimate of the precision matrix. In any case, the normal scores covariance matrix can be computed blazing fast (see our Table 1), and one would expect *CovScores* to be rather robust against cellwise outliers but not so much against casewise outliers.

In order to have at least some chance of dealing with casewise outliers we can transform the data as in (2) and then apply DetMCD instead of the classical covariance. Let us call this *DetMcdScores*. DetMCD is of course much faster than GSE, but since we don't take out cellwise outliers first we would not expect *DetMcdScores* to work as well as 2SGS.

Our Figure 1 shows the performance of 2SGS, *CovScores* and *DetMcdScores* in the case of cellwise outliers (ICM). The clean data were generated according to (1) in dimension p =20 and 40, and then 5% or 10% of the cells were replaced by the values k and $-k$. Since we are not particularly interested in relative scale factors, all scatter matrices were normalized to shape matrices (i.e. their determinant was made 1) before computing their Kullback-Leibler divergence LRT. Naturally 2SGS performs best, whereas the LRT of *CovScores* is higher but at least stays bounded, unlike most other estimators in Figure 1 of ALYZ. *DetMcdScores* is only slightly better here.

For casewise outliers (THCM) the situation changes, as seen in our Figure 2. Now the LRT of *CovScores* grows without bound, whereas that of *DetMcdScores* increases at first but then comes down again, just like 2SGS. Still *DetMcdScores* is not competitive with 2SGS, but it is much faster as seen in our Table 1. This makes us think that it might be worthwhile to invest in creating a faster version of the GSE estimator in order to speed up 2SGS.

Our final comment is about the application in Section 5. There ALYZ compare cellwise distances with the threshold $(\chi^2_1)^{-1}(0.99^{1/(np)})$ and do something similar to the cutoff for the pairwise distances and for the p-wise distances, in order to account for multiple comparison. We don't think that is appropriate here, as the question is to estimate the proportion of outlying cellwise (or pairwise or casewise) distances, as reported in their Table 3. The multiple testing correction would be more justifiable if the question was whether there are any (cellwise, pairwise, or casewise) outliers in the data. Intuitively, if these were not real but generated data, the adjusted formulas would yield quite different proportions in Table 3 merely by doubling the sample size. Of course, comparing with the usual cutoffs like $(\chi^2_1)^{-1}(0.99)$, $(\chi^2_2)^{-1}(0.99)$ and $(\chi^2_p)^{-1}(0.99)$ would yield higher proportions in

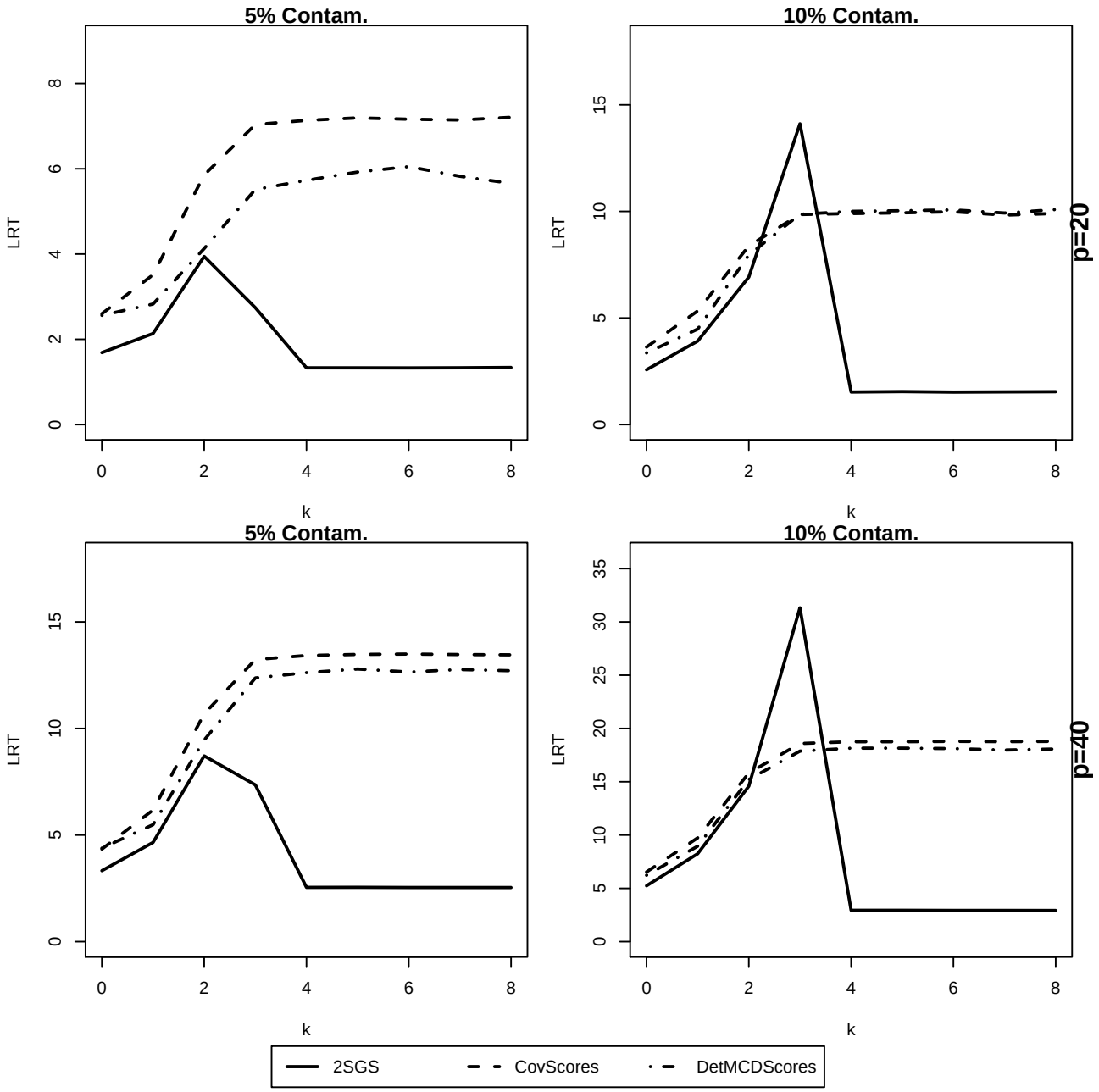

Figure 1: Kullback-Leibler divergence of the 2SGS, CovScores, and DetMcdScores shape matrices under cellwise contamination. The clean data were generated from (1) with $\rho = 0.9$ and then contaminated by 5% and 10% of outliers. Top: 20 dimensions, bottom: 40 dimensions.

Table 3, especially for the casewise outliers, but this is in accordance with the formula $\overline{\varepsilon} = 1 - (1 - \varepsilon)^p$ given by ALYZ which says that ICM yields a lot of contaminated cases.

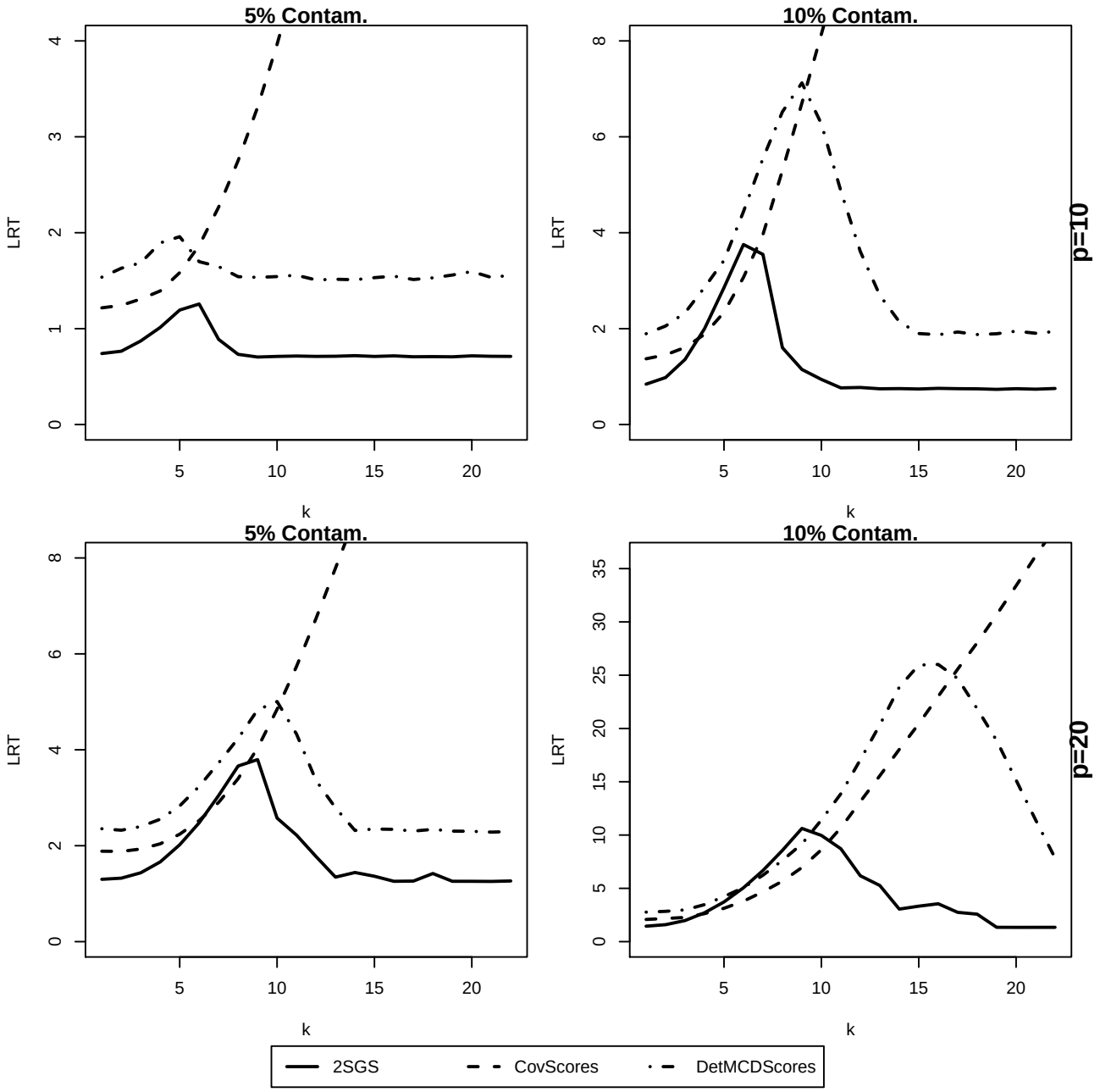

Figure 2: Kullback-Leibler divergence of the 2SGS, CovScores, and DetMcdScores shape matrices under casewise contamination. The clean data were generated as in Figure 1, and then contaminated by 5% and 10% of casewise outliers in the direction of the last eigenvector.

References

Friedman, J., Hastie, T. and Tibshirani, R. (2008), 'Sparse inverse covariance estimation with the graphical lasso', *Biostatistics* **9**, 432–441.

Hubert, M., Rousseeuw, P.J. and Verdonck, T. (2012), 'A deterministic algorithm for robust location and scatter', *Journal of Computational and Graphical*

Statistics **21**, 618–637.

Öllerer, V., Alfons, A. and Croux, C. (2013), 'The shooting S-estimator for robust regression', Technical Report, KU Leuven, Belgium.

Öllerer, V. and Croux, C. (2014), 'Robust high-dimensional precision matrix estimation', Technical Report, KU Leuven, Belgium.